2014-05

iCOP: Automatically Identifying New Child Abuse Media in P2P Networks

Peersman, C

10.1109/SPW.2014.27
2014 IEEE Security and Privacy Workshops
IEEE

iCOP: Automatically Identifying New Child Abuse Media in P2P Networks

Claudia Peersman*, Christian Schulze†, Awais Rashid*, Margaret Brennan‡ and Carl Fischer*
*Security Lancaster Research Centre
Lancaster University, Lancaster, UK
Email: {c.peersman, a.rashid} @lancaster.ac.uk
†German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
Email: christian.schulze@dfki.de
‡School of Applied Psychology
University College Cork, Cork, Ireland
Email: m.brennan@ucc.ie

***Abstract*—The increasing levels of child sex abuse (CSA) media being shared in peer-to-peer (P2P) networks pose a significant challenge for law enforcement agencies. Although a number of P2P monitoring tools to detect offender activity in such networks exist, they typically rely on hash value databases of known CSA media. Such an approach cannot detect new or previously unknown media being shared. Conversely, identifying such new/previously unknown media is a priority for law enforcement – they can be indicators of recent or on-going child abuse. Furthermore, originators of such media can be hands-on abusers and their apprehension can safeguard children from further abuse. The sheer volume of activity on P2P networks, however, makes manual detection virtually infeasible. In this paper, we present a novel approach that combines sophisticated filename and media analysis techniques to automatically flag new/previously unseen CSA media to investigators. The approach has been implemented into the iCOP toolkit. Our evaluation on real case data shows high degrees of accuracy while hands-on trials with law enforcement officers highlight iCOP's usability and its complementarity to existing investigative workflows.**

***Index Terms*—cyber crime; child protection; paedophilia; peer-to-peer computing; image classification; text analysis**

I. Introduction

The proliferation of the Internet and peer-to-peer file sharing systems has transformed the distribution of child sexual abuse media into a crime without geographical boundaries. While there is scientific debate [1] on whether the online paedophile is a new type of offender [2] or if those with a pre-disposition to offend are responding to the opportunities afforded by the new forms of social media [3], empirical evidence points to the problem of Internet-based paedophilia as endemic. Recently, [4] found that 1.6% of searches and 2.4% of responses on the Gnutella P2P network related to illegal sexual content (e.g., rape, bestiality, child abuse). A similar study of the eDonkey network [5] showed that approximately 2 out of every 1,000 users pursued CSA content. Given the networks' scale, these results suggest that, on such file-sharing networks, hundreds of searches for such illegal images occur each second.

The severity of the problem has resulted in a number of solutions that can monitor such activity. Tools such as the Child Protection System (CPS) [6] and RoundUp [7], [8] are able to capture data about paedophile activity on P2P networks and identify child abuse media across different P2P protocols. However, these systems rely on matching the files shared on a network against a hash-value database of known CSA media[1]. As a result, they are not able to identify new child abuse media that may be released on to the network. Nor are they able to detect CSA media that is not on record. Identifying such new/previously unknown media is, however, critical, because they can be indicators of recent or even on-going child abuse. Furthermore, originators of such media can be hands-on abusers and their early detection and apprehension can safeguard their victims from further abuse.

Detecting new/previously unknown CSA media requires (semi-)automatic analysis of image and video content. Though such techniques have already been developed (see e.g. [9], [10]), so far, they have only provided moderate detection rates, due to utilizing single feature descriptions. Also, these tools rely on skin detection techniques, which have been shown to be only marginally discriminative for CSA content detection [11]. Moreover, downloading all files being shared in order to apply such media analysis techniques is clearly infeasible in a P2P scenario.

In this paper, we present a novel approach for automatically detecting and classifying new or previously unknown CSA media on P2P networks. The key contributions of our work are as follows:

- A new filename classification approach that utilises a combination of character *n*-grams and specialised vocabulary used to share CSA media on P2P networks to automatically identify potential candidates for new CSA media out of the millions of files that are being shared.
- An improved image and video classification technique using multiple and, in case of video, multimodal (visual and audio) feature descriptions leading to a robust and

[1]Such databases are built over time through post-hoc forensic analysis of seized computers of offenders.

DOI 10.1109/SPW.2014.27

highly accurate identification of CSA media content.
- A synthesis of the above two analyses, while disregarding files with known hash values, to flag the most pertinent candidates for new/previously unknown child abuse media.
- Operationalisation of the above analyses and synthesis into the iCOP toolkit for use in child protection investigations.

Additionally, we describe the results of evaluating our approach on real CSA filenames and features of CSA media, which show high degrees of accuracy. Furthermore, a user evaluation by law enforcement officers highlights the usability of the iCOP toolkit and its potential to complement and enhance extant investigative workflows pertaining to CSA media.

The rest of this paper is structured as follows. Section II provides an overview of the related work on detecting CSA media and associated activity in P2P networks. Section III describes the key components of our approach, that is, the filename classification and media classification. Section IV, outlines the architecture of the iCOP toolkit and discusses how the two analyses are synthesised to detect new/previously unknown CSA media. Section V presents the evaluation of our approach on CSA data and insights from a user trial. Finally, Section VI concludes the paper, discusses limitations of our approach and identifies directions for future research.

II. Related Work

A. Policing CSA Media on P2P Networks

Our semi-structured interviews with a sample of P2P investigators combined with a survey of law enforcement user requirements exposed a fundamental need for the development of approaches that enable P2P investigators to identify and prioritise cases where the target is engaged in the sexual abuse of children and/or the production of CSA media. Contemporary initiatives to police the exchange of CSA media in P2P networks are frequently played out in operational settings that are bound by scant investigative resources. Our analysis shows how these conditions have exposed a latent conflict of interest between policing initiatives, which give primacy to the identification of victims and those that centre upon the apprehension of the offender, especially the "low hanging fruit" of the offender population whose offending is limited to second-order possession or distribution offences [12]. Whether charged with enforcing the law in respect of broader offences of possession and distribution, or with the apprehension of producers of child abuse media, the identification of contact sexual abuse and abuse victims were cited as paramount concerns for P2P investigators. This finding resonates with earlier observations that a primary goal of P2P investigations is to catch child abusers and help children that are being sexually victimised, rather than simply detecting and confiscating images in the context of possession offences [7], [8]. However noble, these objectives are difficult, nigh impossible to realise using state-of-the-art tools such as CPS and RoundUp. Such tools, which identify suspects involved in the exchange of known CSA files, yield many potential targets for law enforcement but offer little support for the identification and prioritisation of high-risk targets. While some preliminary attempts have been made to utilise materials accessed by suspects to assist in prioritising which investigations take place first (e.g. [13]), no frameworks exist to reliably discriminate high-risk targets in P2P policing contexts – such as those distributing new/previously unknown CSA media that may indicate recent or on-going child abuse.

B. Text Categorisation

Due to the increased availability of documents in digital form and, consequently, the need to access them in quick and flexible ways, content-based document management tasks have acquired a prominent position in the information systems field. One such task consists of automatically labelling natural language texts with a number of predefined thematic categories, i.e. automatic text categorisation, which is currently being used in many different contexts, ranging from document filtering (e.g., spam detection), topic detection and word sense disambiguation to the population of hierarchical catalogues of Internet resources. In current text categorisation studies, the dominant approach to this problem is based on Machine Learning techniques, in which an inductive process automatically builds a classifier by learning the characteristics of the individual categories from a set of documents that were labelled in pre-processing. The trained classifier can, then, distinguish between these categories when it is confronted with new texts showing similar characteristics[2].

Although recent work has tackled the problem of detecting deception [15], masquerading behaviour [16] and identifying paedophile grooming activities [17] online by combining natural language analysis with Machine Learning, these studies all operated on larger bodies of text (e.g. chat room conversations). Contrary to these studies, automatic filename categorisation typically involves much shorter text fragments, which inevitably leads to highly sparse data. Additionally, research has shown that distributors of CSA media tend to use a specialised vocabulary containing a whole variety of abbreviations, acronyms and even combinations of different languages to avoid (automatic) detection of their shared files, while making them widely searchable for other offenders (e.g., "kinderficker", "kdquality", "ptsc") [5]. This vocabulary also proved to be dynamic, i.e., it evolves as existing keywords come to the attention of law enforcement [18].

While a number of studies have focused on textual features that are related to web-accessible images (e.g. [19]–[21]), they do not address any of these additional challenges. So far, there are only two studies – to our knowledge – that used language analysis techniques to identify CSA media. Firstly, [5] investigated the feasibility to automatically construct lists of potential paedopile keywords. Secondly, inspired by previous work on SMS normalisation (see [22], [23] examined whether

[2]An introduction to computational methods for text categorisation can be found in e.g. [14].

these techniques could also be used to circumvent the issue of language variation in CSA filenames. Although their work on pornographic versus non-pornographic filename classification showed very promising results, we found both their approach and the keyword approach [5] ineffective when dealing with the numerous spelling variations incorporated into real CSA filenames.

In contrast to the above approaches, our approach is not only able to deal with the specialised vocabulary used for CSA media, but also takes into account spelling variations and other noise often used to obfuscate CSA filenames in P2P networks.

C. Media Classification

In recent years, quite some work on detecting pornography in images and videos has been published. Most of these studies focus on the utilisation of skin based features (e.g. [24], [25] and [26]) or bag-of-visual-word features (e.g. [27]). These techniques can also be applied to video data by drawing keyframes from the video stream and extracting image features like colour histograms and skin features (e.g. [28]) or skin area shapes (e.g. [29]). For video data, prior research also incorporated acoustic features, such as MFCCs [30] or so-called *audio words* [31], which entails an analogue approach to the visual words technique, based on vector quantisation. Furthermore, motion features, such as motion histograms [32] and the autocorrelation of motion signals [33], have been considered for pornography detection as well.

So far, only few studies have investigated the feasibility to identify CSA data based on visual features (e.g. [34]). The authors of [35] use RGB based skin detection in conjunction with filename analysis and a hash-based detection method. Another quite common approach, especially in forensics, is searching for visually similar image media in an index of already known images as presented in [36] and [37]. However, these methods appear inadequate for the detection of new/previously unknown media.

In our approach, we consider representatives of all of these feature types, except for the motion modality, because they have shown a strong dependency on the utilised video codec and by that generalise poorly for other media sources (see [38]). Contrary to previous approaches to detect CSA media, content describing visual and acoustic features are not used in isolation, but in combination. This has shown promising results for pornography detection in, e.g. [38]. Moreover, instead of using only a single visual modality, we combine a range of various visual and non-visual features for detecting CSA content. Fusing the classification results of different features and modalities enables us to create the most comprehensive trainable CSA content detection system for image and video data to date.

III. Approach

In this section, we elaborate on each of the key components of the iCOP toolkit individually. More specifically, we describe our data collection and preprocessing, together with the feature engineering process. We discuss how the filename and media classification modules are integrated into the architecture of the iCOP toolkit to flag the most pertinent candidates for new/previously unknown child abuse media in Section IV.

A. Filename Classification

Our approach to identifying filenames that possibly contain CSA media is based on automatic text categorisation (see Section II-B). This involves (1) the compilation of a pre-labelled dataset (in our case CSA versus non-CSA filenames), (2) selecting potentially discriminative linguistic features and (3) building a classification model that can attribute new texts automatically to one of the predefined classes with sufficient reliability. However, building a filename classifier that is sufficiently robust so it can be employed by an automatic environment such as the iCOP toolkit is a difficult task for a variety of reasons. Firstly, crawling for CSA files directly from a P2P network to acquire training data is illegal. Hence, we could only use CSA-related filenames that we were allowed to collect from closed court case files. This resulted in a corpus of 268 filenames for the CSA class. Since we aimed to create a classifier that focuses on exactly those textual features that distinguish CSA from adult pornographic media, for the non-CSA class, we crawled filenames that were linked to legal pornography media. Hence, we collected a total of 10,000 non-CSA filenames from *PicHunter*, *PornoHub*, *RedTube* and *Xvideos*[3] (see also [23]). During training, this class imbalance was maintained to simulate a real-life data distribution in a P2P network. Moreover, while prior research mainly focused on automatically identifying and/or normalising typical keywords that are used by Internet paedophiles to camouflage their files' illegal content (see Section II-B), we applied a more comprehensive approach during the feature selection process by combining paedophile keyword information with other linguistic features. More specifically, we first created a dictionary-based filter containing a manually extended version of the paedophile keyword lists from the MAPAP project [5]. We further extended this filter with forms of explicit language use (e.g., "handjob") and expressions relating to children (e.g., "kiddie") and family relations (e.g., "daughter"). Together, these four categories, i.e., the *paedophile keywords*, the *explicit keywords*, the *child references* and the *family references*, form our **Semantic features**. Hence, a filename without any paedophile keywords can still become a high-value target with regard to CSA media when it contains, for example, both explicit keywords and references to children. We show an example of the feature construction in Table I. The presence of the keyword "pt" (*preteen*) results in a hit for the paedophile keywords category, while "12yo" (*12 years old*) is identified as a reference to a child.

Secondly, we extracted all patterns of two, three and four consecutive characters from the filenames (also called **Character *n*-gram features**). As can be seen from the example in Table I, this approach allowed us to circumvent the issue

[3]www.pichunter.com, www.porno-hub.com, www.redtube.com, www.xvideos.com

of alternative keyword spellings: although the actual keyword "lolita" is not present in the example filename, the presence of the "lita" feature could be equally discriminative when training the classifier, because that feature is also present in filenames that do contain the original keyword. Additionally, other potential cues could be detected by the model, even when they were related to new/unknown keywords that are not (yet) included in the semantic features.

TABLE I
EXAMPLE OF A CSA FILENAME AFTER FEATURE CONSTRUCTION

Original filename	ptl0lita12yo.jpeg
2-gram feats.	pt tl l0 0l li it ta a1 12 2y yo
3-gram feats.	ptl tl0 l0l 0li lit ita ta1 a12 12y 2yo
4-gram feats.	ptl0 tl0l l0li 0lit lita ita1 ta12 a12y 12yo
Semantic feats.	paedo_keyword child_ref

B. Media Classification

As we process both images and videos, the content classification module contains two input streams: (a) images, being fed into the feature extraction pipeline directly, and (b) video files, that are pre-processed by extracting video frames and a continuous audio stream. For frame extraction, the input video is split into shots of 100 frames. The centre frame of each of these video segments is taken as a representative keyframe for the extraction of visual features. Additionally, audio features are computed for all $4s$ segments, respectively. For describing the visual content of images and video frames we extract: colour-correlograms, skin features, visual words and visual pyramids. The audio information of video files is described by computing vector quantised MFCC features (i.e. *audio words*). In the following, a brief description of the utilised feature extractions is presented:

- **Colour-Correlograms** describe the occurrence probability of a colour in a pixel's neighbourhood (see e.g. [39], [40], [41], [42]). Hence, they represent the local spatial correlation of colours in images. Here, we apply a special variant of the colour-correlogram, namely the *auto-colour-correlogram*, which describes the probability of the identical colour c reoccurring within a distance d of the current pixel in image I.

$$\alpha_c^d(I) = \gamma_{c,c}^d(I) \tag{1}$$

  For an optimal performance, this feature is computed in HSV colour space ([40]).
- The **Skin-Feature** is based on a RGB skin colour model, which was generated using manually segmented images from the COMPAQ database. The presence of skin is indicated via a *skin-probability-map* (SPM) (see also [25]).

$$P(skin|c) = \frac{P_{skin}(c)}{P_{skin}(c) + P_{non-skin}(c)} \tag{2}$$

  For computing the skin feature, first the SPM is transferred into a *skin-segmentation-mask* (SSM) via morphological operations and adaptive thresholding. Next, the mean intensities of SPM and SSM are calculated, as well as their center and variance of skin mass. This yields a 14 dimensional descriptor representing image skin properties.
- Because colour features – especially skin features – are not very robust towards illumination changes, we also computed **Visual Words and Pyramids** to provide a texture based content representation. Visual words features are computed by scaling the given image to 250×250 and extracting patches of 8×8 using a regular sampling with a step size of 5 pixels. Next, the DCT coefficients of the YUV transformed patches are computed for each channel. The final feature is represented by 36 low frequency coefficients from the Y-channel and 21 taken from U and V, respectively. Our visual pyramid features are based on the same representation and are structured according to [44]. Applying $2,000$ entry codebooks for vector quantisation yields the final representation of both features.
- The audio stream of video files is described by extracting **Audio Words**, using the widely used Mel Frequency Cepstral Coefficients (MFCC) [43]. We extract MFCCs in steps of $8ms$, using a $16ms$ sliding window. Next, a frequency histogram is computed from the Fourier transform of the signal. For reflecting human acoustic perception, the frequency histogram is weighted by the logarithmic Mel scale. Finally, the weighted histograms are DCT encoded, leading to a 13 dimensional descriptor, which are vector quantised using a $1,000$ entry codebook.

IV. THE ICOP TOOLKIT

The filename and image classification approaches are synthesised in the iCOP toolkit to identify new/previously unknown CSA media. As shown in Figure 1, the toolkit has two major components: the P2P Engine and the iCOP Analysis Engine.

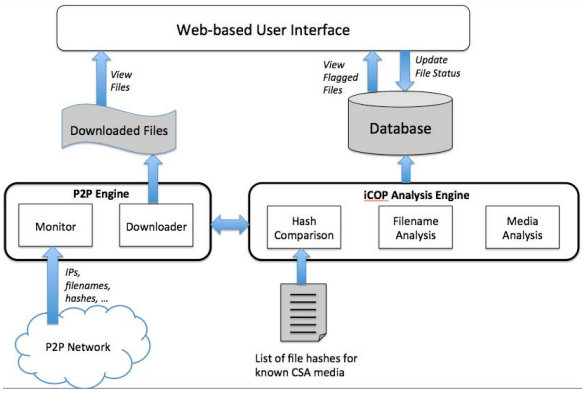

Fig. 1. Overview of the iCOP toolkit

The P2P engine provides functionality to monitor public traffic on a P2P network. Currently, the P2P engine supports monitoring of such traffic on Gnutella but other monitors can be plugged into the engine. The monitor extracts information such as IP addresses, filenames and hash values of files as well as meta data, such as when a particular peer was last seen sharing a file. The latter is essential to identify originator of a new CSA file. This information is passed on to the iCOP analysis engine which undertakes the following steps:

1) It compares the hash values of files against a list of known hashes. As we mentioned above, such hash value lists are established by law enforcement when CSA media are seized. This filtering mechanism ensures that the system disregards known CSA media. The user interface does indicate when a peer is sharing known CSA media, but the toolkit does not download or process the files given the focus on identifying new/previously unknown CSA media. This significantly reduces both storage and computation requirements. We currently use a file of SHA1 hashes in base-32 (one hash per line). We do so because this is the most common format in which law enforcement store hash values for CSA media. The design enables law enforcement officers using the toolkit to plug-in their own hash value lists without substantial effort to import them into a specific format or database.
2) The names of files that do not occur in the known hash list are passed on to the filename classifier for identifying their likelihood of containing CSA media. File names that are deemed to be non-CSA media are discarded.
3) Files that are flagged by the filename classifier as potentially containing CSA media are passed back to the P2P engine for downloading. The downloaded files are piped back to the iCOP analysis engine, more specifically, to the media classifier to determine if the content is indeed CSA media.

The results of the analysis are stored in real-time in the database. An investigator can login to the GUI to access the iCOP "dashboard" which flags the most pertinent candidates for CSA media as the highest priority. Additionally, the user can view thumbnails as well as the full media files to verify whether the flagged items are indeed CSA media. If so, these items can be marked "confirmed" by the user and are fed back into the hash database so that they are considered to be known child abuse media in future searches. This setup enables the toolkit to triage the files that need to be downloaded and analysed by the image classifier.

The toolkit GUI is designed around a list of connections, which maps closely to the way P2P software works. A connection is defined as:

connection = IP address + Port + GUID

Each connection is assumed to be a single user sharing a given set of files from a specific location. This is in contrast with an IP address alone, which could potentially be shared by multiple users (e.g., several machines in a home) or a GUID alone, which could potentially be used from different locations (e.g., work, home, travel). The toolkit can display files shared by a particular IP or a particular GUID. Hence, an investigator can easily view which connections are related via a common IP address or GUID. As mentioned above, the most pertinent candidates are flagged to the user as high priority via the dashboard. Given legal constraints governing law enforcement, the toolkit can also be configured to focus on particular geo-locations (e.g., a particular country or region). Additionally, the toolkit provides a demo mode to allow testing and debugging the toolkit using dummy P2P network data and legal pornography. This is because any monitoring and downloading of CSA media can only take place at suitable law enforcement premises.

V. Evaluation

A. Filename Classification

To obtain a reliable estimation of the classifier's performance, we performed five-fold cross validation (see [45]) during the experiments. In this experimental regime, the available data is randomised and divided into five equally sized folds or partitions. Subsequently, each partition is used four times in training and once in test. For classification, we used the SVM algorithm as implemented in LibShortText [46], an open-source software package for short-text classification and analysis. Parameters were experimentally determined on a development set of each training partition during cross validation. The scores we report are average *precision*, *recall* and *F-score*. These are standard evaluation metrics that can be computed based on the number of true positives (*tp*), true negatives (*tn*), false positives (*fp*) and false negatives (*fn*) in a confusion matrix. The recall score for each class provides information on the number of filenames that were successfully retrieved, while the precision score takes into account all retrieved filenames for each class and evaluates how many of them were actually relevant. The F-score is then the harmonic mean of precision and recall. These measures are defined as follows.

$$Precision = \frac{tp}{tp + fp} \tag{3}$$

$$Recall = \frac{tp}{tp + fn} \tag{4}$$

$$F_{score} = 2\,\frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

The results of the experiments are shown in Table II. The best results were achieved when combining the character *n*-gram features with the semantic features, leading to an overall precision of 89.9%. We further tested our trained classifier with a sample of 40,000 CSA filenames and 40,000 legal pornographic filenames. This resulted in a 92.9% precision, a 52.5% recall and an F-score of 67.1% for the CSA filenames and an overall accuracy of 73.0%.

TABLE II
RESULTS OF THE FILENAME CLASSIFIER USING DIFFERENT FEATURE TYPES

Scores (%)	Precision	Recall	F-score
CSA FILENAMES			
Semantic feats.	5.7	21.3	9.0
Char. n-grams	89.8	62.3	73.6
Combined	**89.9**	**66.1**	**76.1**
NON-CSA FILENAMES			
Semantic feats.	97.7	90.6	94.0
Char. n-grams	99.0	**99.8**	99.4
Combined	**99.1**	**99.8**	**99.5**

B. Media Classification

For our experiments in classifying CSA media, we decided to evaluate two scenarios that have been triggered by practical aspects of law enforcement investigations: (1) detection of CSA content versus regular media content (world) and (2) separation of CSA from legal pornographic media (adult). As CSA media content can be considered a subclass of pornography, the second scenario is expected to be much more challenging. All three content classes were represented by $20,000$ images and $1,000$ short videos each, which have been collected from various web sources like *flickr.com, youtube.com, pichunter.com, redtube.com, and pornhub.com*. Numerical feature representations for instances of CSA media were provided by European law enforcement. During the experiments, first, we extracted all features from the data (see Section III-B), followed by a selection of $1,500$ training and $3,000$ test samples representing positive and negative classes in equal amounts. Next, all extracted features f were presented to separate statistical classifiers. We use SVM's, as they have shown superior performance compared to other options (e.g. [38]). For estimating the SVM parameters C and γ, we performed a 5-fold cross validation. Finally, the scores of the individual classifiers were combined using a weighted sum *late fusion* scheme, yielding a multi-modal classification score for a $4s$ video segment or image X.

$$P(CSA|X) = \sum_f w_f \cdot P^f(CSA|X) \tag{6}$$

The weights w_f for late fusing the trained classifiers were found by grid searching possible classifier combinations. Averaging the performance of all 5 folds provided the numerical results of the experiments, presented in terms of *average precision* (AvP) and *equal error rates* (EER). Both measures are commonly used in information retrieval and biometrics, whereas the AvP is based on a ranked list of results and represents the area under the recall precision curve. It can be computed according to Equation 7, with r being the current rank of the sorted list of N items, $P(r)$ the precision at rank r, $rel(r)$ an indicating function being 1 if the item at r is relevant and 0 otherwise, and D_r the number of relevant items in the analysed set.

$$AvP = \frac{\sum_{r=1}^{N} P(r) \times rel(r)}{D_r} \tag{7}$$

Instead, the EER marks the rate at which the number of false positive and false negative classifications are equal.

As can be seen in Tables III and IV, our classifiers reach average precision in excess of 92% (image) and 95% (video) when compared with adult pornography.

TABLE III
LATE FUSION WEIGHTS w_f AND CLASSIFICATION RESULTS (SINGLE AND FUSED) FOR CSA DETECTION IN IMAGES.

	CSA vs World			CSA vs Adult		
Feature	w_f	**AvP**	**EER**	w_f	**AvP**	**EER**
ccorrelogram	0.55	**92.93**	**14.03**	0.65	**91.07**	**16.83**
vispyramids	0.40	91.36	16.08	0.35	87.41	20.58
skin segment	0.05	81.32	26.43	0.00	74.24	33.60
fused		**94.69**	**11.70**		**92.09**	**15.53**

TABLE IV
LATE FUSION WEIGHTS w_f AND CLASSIFICATION RESULTS (SINGLE AND FUSED) FOR CSA DETECTION IN VIDEOS.

	CSA vs World			CSA vs Adult		
Feature	w_f	**AvP**	**EER**	w_f	**AvP**	**EER**
audiowords	0.40	88.34	16.23	0.45	**90.24**	**15.28**
ccorrelogram	0.30	90.15	13.95	0.35	86.07	16.08
vispyramids	0.20	89.88	14.08	0.20	82.02	19.44
viswords	0.10	**90.38**	**13.77**	0.00	79.41	20.45
fused		**97.32**	**7.51**		**95.65**	**8.19**

C. User Evaluation

We conducted a live testing workshop with 9 law enforcement officers engaged in CSA investigations on P2P networks from 7 different law enforcement agencies across Europe. Given the legal (as well as ethical) issues pertaining to such a live exercise, the two day workshop was conducted on law enforcement premises. The participants were provided background information on the toolkit as well as details of how the analysis is performed by the backend. They were provided training in the use of the user interface followed by actual use of the toolkit on live data. A lot of usability feedback was gathered in focus-group style discussions and used to improve the functionality and the user interface subsequently. At the end of the workshop, a questionnaire was completed by the participants. Each question was answered according to a 7-point Likert scale (1 - strongly disagree ... 7 - strongly agree) and had room for comments. The questions were as follows:

1) Overall, the toolkit is easy to use.
2) It was easy to learn to use this toolkit.
3) The toolkit has all the capabilities I need to prioritise investigations.
4) I believe the toolkit can facilitate my investigations.

5) I believe the toolkit will allow me to more efficiently carry out investigations.
6) I believe the toolkit will assist me in efficiently analysing the large number of files shared on P2P networks.
7) I would frequently use this toolkit as part of my investigations.
8) Overall, I believe the toolkit to be a valuable aid to law enforcement.

The results are summarised in Table V where Q is the question and P the participant. Participant 8 was a consistent outlier in terms of low scores. Hence, we present the results after omitting the outlier responses from P8. Particularly noteworthy is the positive feedback for ease of use (Q1, 5.4/7), ease of learning (Q2, 5.8/7), facilitation of investigations (Q4, 5.8/7), and value for law enforcement (Q8, 6.1/7).

TABLE V
SURVEY RESULTS

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Mean	Std
P1	7.0	7.0	6.0	6.0	6.0	5.0	5.0	6.0	6.0	0.7
P2	6.0	6.0	6.0	6.0	6.0	5.0	5.0	6.0	5.8	0.4
P3	7.0	7.0	4.5	6.0	6.0	4.0	7.0	7.0	6.1	1.1
P4	3.0	4.0	4.0	4.0	3.0	6.0	5.0	4.0	4.1	0.9
P5	6.0	6.0	6.0	7.0	7.0	5.0	6.0	7.0	6.3	0.7
P6	6.0	7.0	5.0	5.0	5.0	4.0	5.0	7.0	5.5	1.0
P7	5.0	5.0	6.0	7.0	6.0	7.0	5.0	7.0	6.0	0.9
P9	3.0	4.0	3.0	5.0	3.0	4.0	3.0	5.0	3.8	0.8
Mean	5.4	5.8	5.1	5.8	5.3	5.0	5.1	6.1	5.4	
Std	1.6	1.3	1.2	1.0	1.5	1.1	1.1	1.1		

VI. CONCLUSION

In this paper, we have presented the iCOP toolkit: a forensic software package that is designed to highlight sharers of new or unknown child sexual abuse media in P2P networks. Additionally, it offers secondary features, such as showing sharers of known CSA files and allowing law enforcement investigators to see other files shared by the same computer or other IP addresses used by the same P2P client. Hence, the software allows law enforcement to more rapidly locate the producers of such content and the victims therein.

Although the current realisation of both the filename and the image classification modules already provide very good results, they could be further optimised. Firstly, the classification of images is still limited by operating only on low-level visual features. Future research can potentially address this in multiple ways. While CSA video classification can be significantly improved by additionally using other modalities, i.e., audio or motion information, image classification could be extended by utilising high-level visual features. For example, the novel SentiBank feature [47], which consists $1,200$ classifier scores indicating the presence of pre-trained concepts, could achieve some orthogonality towards low-level descriptions in feature space, because they build up on different information sources. Another challenge for future research is the age verification of individuals appearing in questioned images and videos. Though evaluations have been conducted during the development of the media classification module, current approaches to determine the age of persons for supporting the classification decision cannot provide the robustness that is needed in an automated environment such as the iCOP toolkit. Also, the filename classification module could be enhanced by retraining on larger datasets of CSA filenames.

Furthermore, the same techniques for monitoring, analysing file names, and analysing content that we propose for the Gnutella network could also be applied to other file sharing systems, such as eDonkey and Bittorrent. Monitoring these networks, however, will require different software libraries for each protocol and may yield different types of information that require slightly different database structures. In addition, some protocols rely on central servers and will need to be manually configured to monitor the servers of interest. Finally, quite a few networks are designed to provide anonymity and prevent freeloading. These issues will provide a further venue for our future work.

ACKNOWLEDGMENT

This work was funded by the European Commission Safer Internet Programme project (SI-2010-TP-2601002), *iCOP: Identifying and Catching Originators in Peer-to-Peer Networks*, the Impact Acceleration Account of Lancaster University and by the Antwerp University, *DAPHNE: Defending Against Paedophiles in Heterogeneous Network Environments*. The authors would also like to thank all law enforcement agencies that have contributed to this project. Finally, special thanks go out to Interpol. Without their efforts, our research would have been impossible.

REFERENCES

[1] D. Middleton, "Internet Sex Offenders", Ch. 12 in Assessment and Treatment of Sex Offenders: A Handbook, 2009.
[2] E. Quayle, G. Holland, C. Linehan, M. Taylor, "The Internet and Offending Behaviour: A Case Study", *Journal of Sexual Aggression*, vol. 6, pp. 78-96, 2000.
[3] A. Cooper, "Sexuality and the Internet: Surfing into the New Millenium", *Cyberpsychology and Behavior*, vol. 1, no. 2, pp. 187-193, 1998.
[4] D. Hughes, J. Walkerdine, G. Coulson, S. Gibson, "Is Deviant Behaviour the Norm on P2P File-Sharing Networks?", *IEEE Distributed Systems Online*, vol. 7, no. 2, 2006.
[5] M. Latapy, C. Magnien, R. Fournier, "Quantifying paedophile activity in a large P2P system", *Information Processing and Management*, vol. 49, no. 1, pp. 248-263, 2013.
[6] Child Protection System. P2P Monitoring software developed at TLO, http://www.tlo.com/, USA.
[7] M. Liberatore, R. Erdely, T. Kerle, B. Levine, C. Shields, "Forensic Investigation of Peer-to-Peer File Sharing Networks", In *Proceedings of the DFRWS Annual Digital Forensics Research Conference*, 2010.
[8] M. Liberatore, B. Levine, C. Shields, "Strengthening forensic investigations of child pornography on P2P networks", In *Proceedings of the ACM Conference on emerging Networking EXperiments and Technologies (CoNEXT)*, 2010.
[9] FIVES: Forensic Image and Video Examination Support. EC Safer Internet project. More information can be found at http://fives.kau.se
[10] I-Dash: The Investigator's Dashboard. EC Safer Internet project. More information can be found at http://www.i-dash.eu/

[11] C. Schulze, D. Henter, D. Borth, A. Dengel, "Automatic Detection of Child Pornography by Multi-modal Feature Fusion for Law Enforcement Support", In *Proceedings of the International Conference on Multimedia Retrieval*, 2014, to appear in.

[12] Y. Jewkes, C. Andrews, "Policing the filth: the problems of investigating online child pornography in England and Wales", *Policing and Society*, vol. 15, no. 1, pp. 42-62, 2005.

[13] M. McManus, M. Long, L. Alison, "Child Pornography Offenders: Towards an evidence-based approach to prioritizing the investigation of indecent image offences", In L. Alison, L. Rainbow (eds.), *Professionalizing offender profiling: Forensic and investigative psychology in practice*, pp. 178-188, 2011.

[14] F. Sebastiani, "Machine learning in automated text categorization", *ACM Computing Surveys*, vol. 34, pp. 147, 2002.

[15] S. Afroz, M. Brennan, R. Greenstadt, "Detecting Hoaxes, Frauds, and Deception in Writing Style Online", In *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 461-475, 2012.

[16] A. Rashid, A. Baron, P. Rayson, C. May-Chahal, P. Greenwood, J. Walkerdine, "Who Am I? Analyzing Digital Personas in Cybercrime Investigations", *IEEE Computer*, vol. 46, no. 4, pp. 54-61, 2013.

[17] G. Inches, F. Crestani,"Overview of the International Sexual Predator Identication Competition at PAN-2012", In P. Forner, J. Karlgren, and C. Womser-Hacker (eds.), *CLEF 2012 Evaluation Labs and Workshop Working Notes Papers*, 2012.

[18] D. Hughes, P. Rayson, J. Walkerdine, K. Lee, P. Greenwood, A. Rashid, C. May-Chahal, M. Brennan, "Supporting Law Enforcement in Digital Communities through Natural Language Analysis", In *Proceedings of the International Workshop on Computational Forensics*, 2008.

[19] F. Wang, M. Kan, "NPIC: Hierarchical synthetic image classification using image search and generic features", In *Proceedings of the Conference on Image and Video Retrieval*, pp. 473-482, 2006.

[20] W. Huang, W. Huang, C. Tan, W. Leow, "Model-based chart image recognition", In *Proceedings of the International Workshop on Graphics Recognition*, pp. 87-99, 2003.

[21] E. Munson, Y. Tsymbalenko, "To search for images on the web, look at the text, then look at the images", In *Proceedings of the 1st international workshop on Web Document Analysis*, http://www.csc.liv.ac.uk/ wda2001, 2001.

[22] R. Beaufort, S. Roekhaut, L. Cougnon, C. Fairon, "A hybrid rule/model-based finite-state framework for normalizing SMS messages", In *Proceedings of ACL*, pp. 770-779, 2010.

[23] A. Panchenko, R. Beaufort, C. Fairon, "Detection of Child Sexual Abuse Media on P2P Networks: Normalization and Classification of Associated Filenames". In *Proceedings of Workshop on Language Resources for Public Security Applications of the 8th International Conference on Language Resources and Evaluation (LREC)*, 2012.

[24] H. Rowley, Y. Jing, S. Baluja. "Large Scale Image-Based Adult-Content Filtering", *Int. Conf. Comp. Vis. Theory and Applications*, pp. 290-296, 2006.

[25] M. Jones, J. Rehg, "Statistical Color Models with Application to Skin Detection", *Int. J. Comput. Vision*, vol. 46, no. 1, pp. 81-96, 2002.

[26] M. Fleck, D. Forsyth, C. Bregler, "Finding Naked People", *ECCV*, pp. 593-602, 1996.

[27] T. Deselaers, L. Pimenidis, H. Ney, "Bag-of-Visual-Words Models for Adult Image Classification and Filtering", *ICPR*, pp. 1-4, 2008.

[28] H. Lee, S. Lee, T. Nam, "Implementation of high performance objectionable video classification system", *ICACT*, pp. 959-962, 2006.

[29] C. Kim, O. Kwon, W. Kim, S. Choi, "Automatic System for Filtering Obscene Video", In *Proceedings of the 10th International Conference on Advanced Communication Technology*, pp. 1435-1438, 2008.

[30] H. Zuo, O. Wu, W. Hu, B. Xu, "Recognition of Blue Movies by Fusion of Audio and Video", In *Proc. of ICME*, pp. 37-40, 2008.

[31] Y. Liu, X. Wang, Y. Zhang, S. Tang, "Fusing Audio-Words with Visual Features for Pornographic Video Detection", In *Proceedings of TRUSTCOM'11*, pp. 1488-1493, 2011.

[32] C. Jansohn, A. Ulges, T. Breuel, "Detecting Pornographic Video Content by Combining Image Features with Motion Information", *ACM Multimedia*, 2009.

[33] T. Xiaofeng, L. Duan, C. Xu, Q. Tian, L. Hanqing, J. Wang, J. Jin, "Periodicity Detection of Local Motion", *IEEE International Conference on Multimedia and Expo*, pp. 650-653, 2005.

[34] A. Ulges, A. Stahl, "Automatic Detection of Child Pornography using Color Visual Words", In *Proc. of the Int. Conf. Multimedia and Expo*, 2011.

[35] P. da Silva Eleuterio, M. de Castro Polastro, "An adaptive sampling strategy for automatic detection of child pornographic videos", *Int. Conf. on Forensic Computer Science*, pp. 12-19, 2012.

[36] LTU Engine, available from http://www.ltutech.com/en/products/ltu-engine-2 (retrieved: June 2010).

[37] Netclean Analyze, available from http://www.netclean.com/ (retrieved: June 2010).

[38] A. Ulges, C. Schulze, D. Borth, A. Stahl, "Pornography detection in video benefits (a lot) from a multi-modal approach", Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis, ACM, 2012.

[39] J. Huang, S. Kumar, M. Mitra, W Zhu, R. Zabih, "Image Indexing Using Color Correlograms", *CVPR*, pp. 762-768, 1997.

[40] T. Ojala, M. Rautiainen, E. Matinmikko, M. Aittola. "Semantic Image Retrieval with HSV Correlograms", In *Proceedings of the 12th Scandinavian Conference on Image Analysis*, pp. 621-627, 2001.

[41] M. Rautiainen, T. Ojala, "Color Correlograms in Image and Video Retrieval", In *Proceedings of the 10th Finnish Artificial Intelligence Conference*, pp. 203-212, 2002.

[42] Z. Zha, Y. Liu, T. Mei, X. Hua, "Video concept detection using support vector machines - trecvid 2007 evaluations", Technical report, 2008.

[43] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling", *Int. Symposium on Music Inf. Retrieval*, 2000.

[44] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp 2169-2178, 2006.

[45] S. Weiss, C. Kulikowski, *Computer Systems that Learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*, Morgan Kaufmann: San Mateo, USA, 1991.

[46] H. Yu, C. Ho, Y. Juan, C. Lin, "LibShortText: A Library for Short-text Classification and Analysis", Technical Report, http://www.csie.ntu.edu.tw/ cjlin/papers/libshorttext.pdf, 2013.

[47] D. Borth, R. Ji, T. Chen, T. Breuel, S. Chang, "Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs", *ACM Multimedia*, 2013.